The Earth System Grid Federation

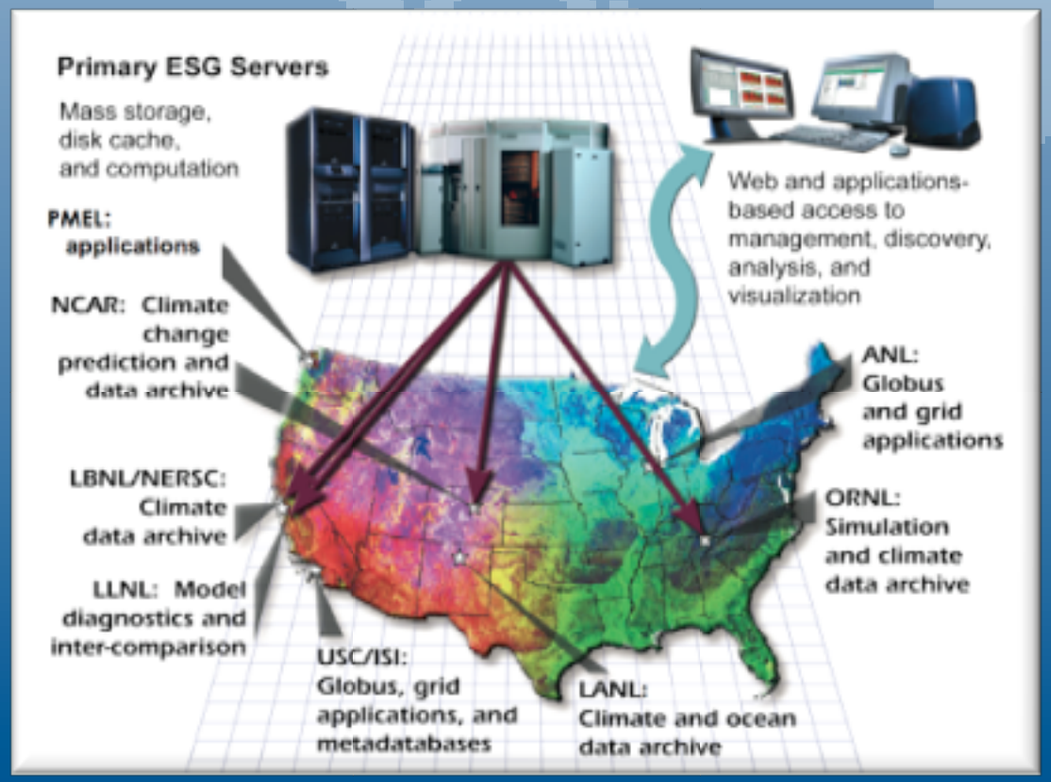
Luca Cinquini [1][2], Dan Crichton [1], Cecelia DeLuca [2], Dean Williams [3]
on behalf of the ESGF collaboration

ESIP Workshop on “Climate and Energy”
Knoxville, TN, July 2010

[1] Jet Propulsion Laboratory (JPL), NASA, operated by the California Institute of Technology
[2] Earth System Research Laboratory (ESRL), NOAA
[3] Program for Climate Model Diagnostic and Inter-comparison (PCMDI), LLNL, funded under the U.S. Department of Energy's Office of Science, Biological and Environmental Research (BER) Program

The Earth System Grid Federation (ESGF)

- ESGF is an open consortium of institutions, laboratories and centers around the world that are dedicated to supporting research of Climate Change, and its environmental and societal impact
- Historically originated from Earth System Grid project, expanded beyond its constituency and mission to include many other partners in U.S., Europe, Asia and Australia
- Groups working at many projects: ESG, ESC, Metafor, GIP, IS-NES...
- U.S. funding from DOE, NASA, NOAA, NSF



- U.S.: PMEL, LLNL/PCMDI, LBNL, USC/ISI, NCAR, LANL, ORNL, ANL, JPL, GFDL, ESRL
- Europe: BADC, UK-MetOffice, DKRZ, MPIM, IPSL, LSCE
- Asia: Univ. of Tokyo, Japanese Centre for Global Environmental Research, Jamstec, Korea Meteorological Administration
- Australia: ANU, Australian Research Collaboration Service, Government Department of Climate Change, Victorian Partnership for Advanced Computing, Australian Environment and Resource Management
- ... and more ...

GO-ESSP

- ESGF governance and coordination is provided by the Global Organization for Earth System Science Portals (GO-ESSP)
- "The Global Organization for Earth System Science Portals (GO-ESSP) is a collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data from the climate and weather communities".
- "Grass root" effort, unofficially started at LLNL in 2001, meetings every 12-18 months
- Very broad focus, even beyond ESGF
  - Climate
  - Weather

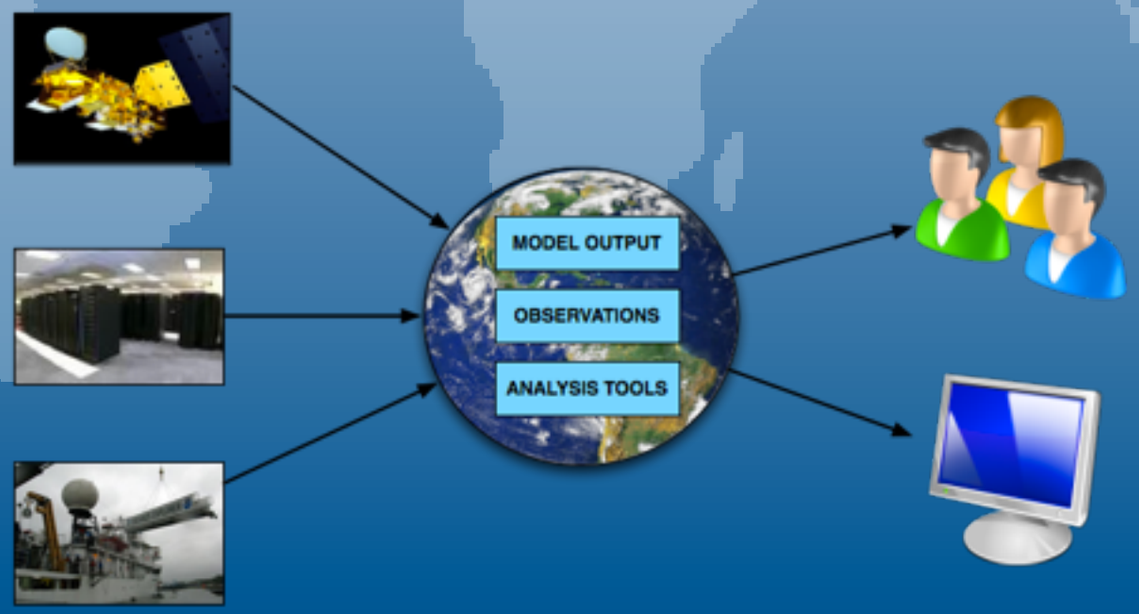
http://go-essp.gfdl.noaa.gov/

ESGF Goals

To build, operate and support a global infrastructure for the management, access and analysis of climate data

- Promote sharing of knowledge, software and tools among partners
- Define APIs and protocols for interoperability among data centers
- Collaborative development of some software components
- Deployment of common software infrastructure

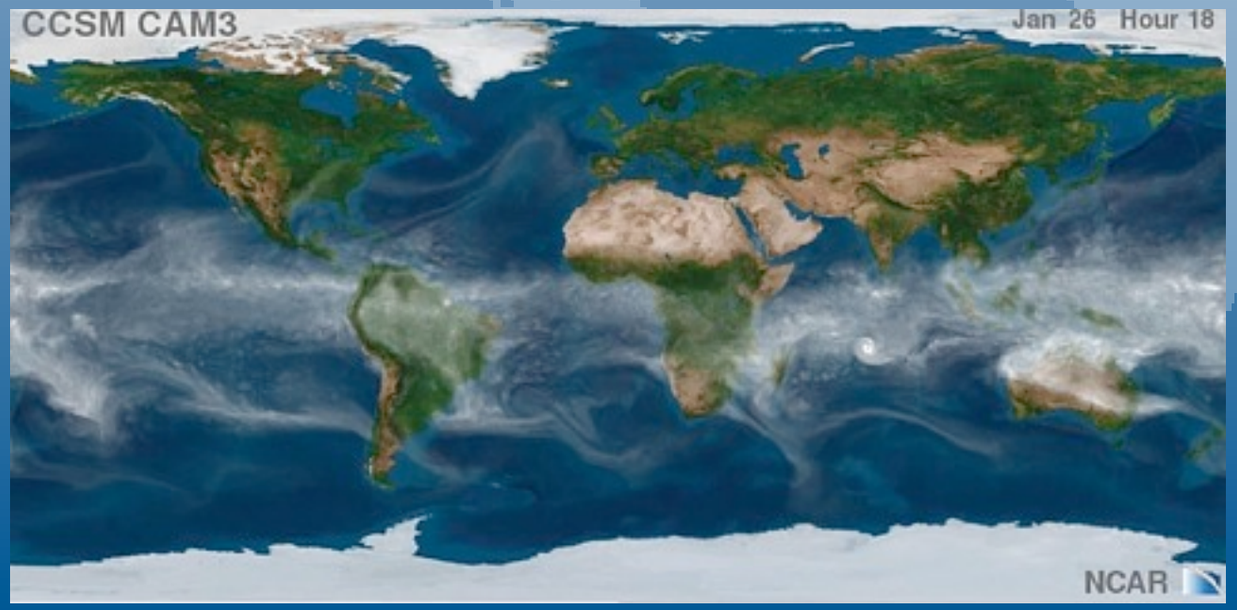
To build a virtual federated environment that combines model output, observational data, analysis & visualization tools, and facilitates and empowers access by all user communities

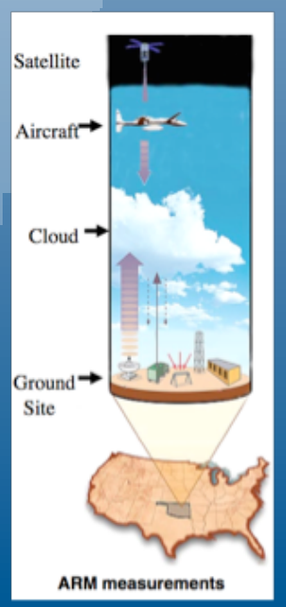


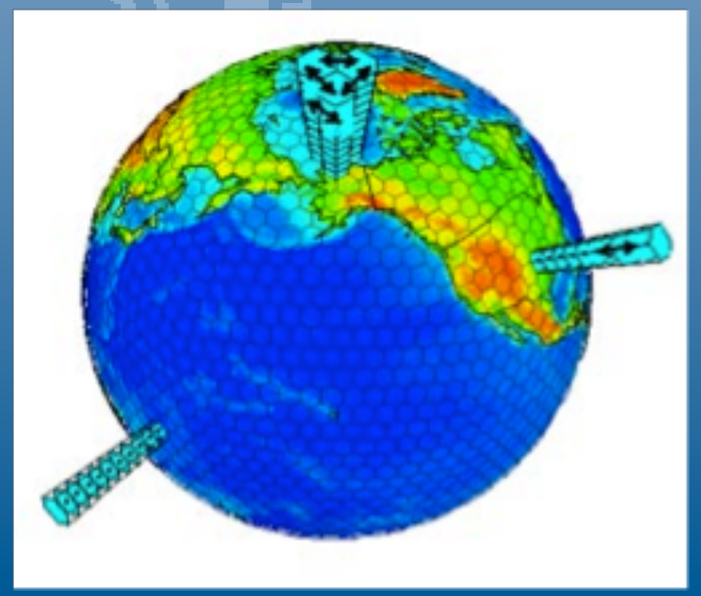
ESGF Challenges

Climate change analysis is a global problem which presents unprecedented challenges from the data management and distribution perspective:

- Massive data archives (many PB moving to XB)
- Multiple data centers world-wide
  - existing IT infrastructures and separate security domains
- Heterogenous data sources (models, observations, reanalysis)
- Multiple data and metadata formats (NetCDF, CF, HDF, HDF-EOS…)
- Multiple physical realms (atmosphere, ocean, land, sea ice)
- Multiple scales (global, regional and local)
- Multiple audiences (scientists, policy makers, students, educators)







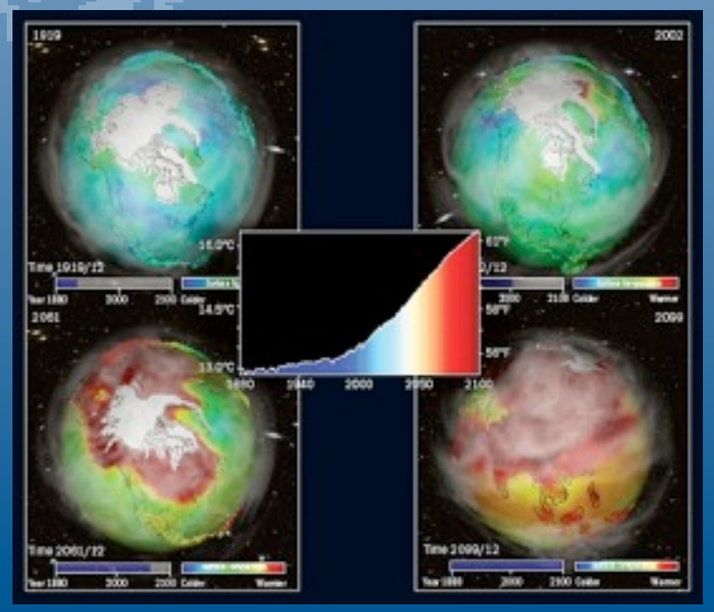
CMIP5 and IPCC-AR5

- CMIP5: Coupled Model Inter-comparison Project, phase 5
  - Activity sponsored by WCRP (World Climate Research Program) to promote study of climate change
  - Global archive of 40+ models contributed by 25+ modeling centers in 17+ countries
  - 3 categories of experiments: “Near-Term” decadal predictions, “Long-Term century & longer”, “Atmosphere Only”
  - 1.2-2PB of replicated “core” data, 10+ PB total data
- IPCC-AR5 (Intergovernmental Panel on Climate Change – 5th Assessment Report) will use “core” data in CMIP5 archive

ESGF role in CMIP5 and IPCC-AR5

- CMIP5/IPCC-AR5 are major drivers for ESGF infrastructure development and collaboration
  - Functionality, data volumes, timescale (2010-2020)
- Role of ESGF:
  - Consortium of 30+ national and international partners that are involved in different roles in the CMIP5 process (modeling centers, archives, data distribution and analysis gateways)
- Tasked with deploying and operating the global infrastructure for CMIP5 data archival and access
- Three Replica Centralized Archives (RCAs)
  - PCMDI: continuos growth with CMIP data
  - BADC and DKRZ: frozen in 2012/13 for AR5

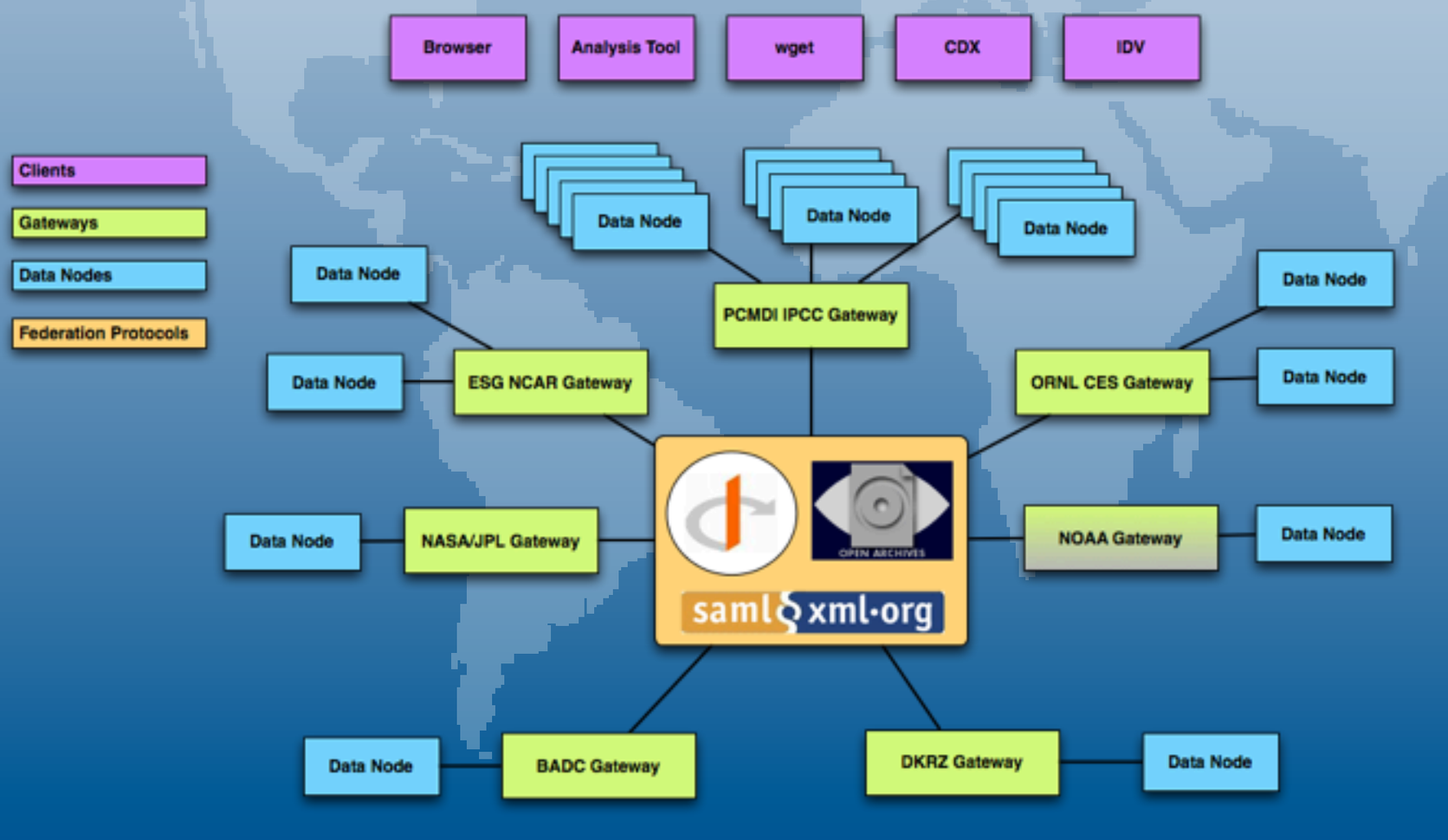
ESGF will play a critical role in facilitating and empowering climate change research at a global level



ESGF High Level Architecture

ESGF system is a network of Gateways and Data Nodes

- Data is stored and accessed from distributed archives
- Institutions maintain autonomous control of resources…
- …but inter-operate through common federation protocols
- Software stack based on integration of services and tools commonly adopted in geo-scientific community
- Enabling access to browsers and rich desktop-based clients

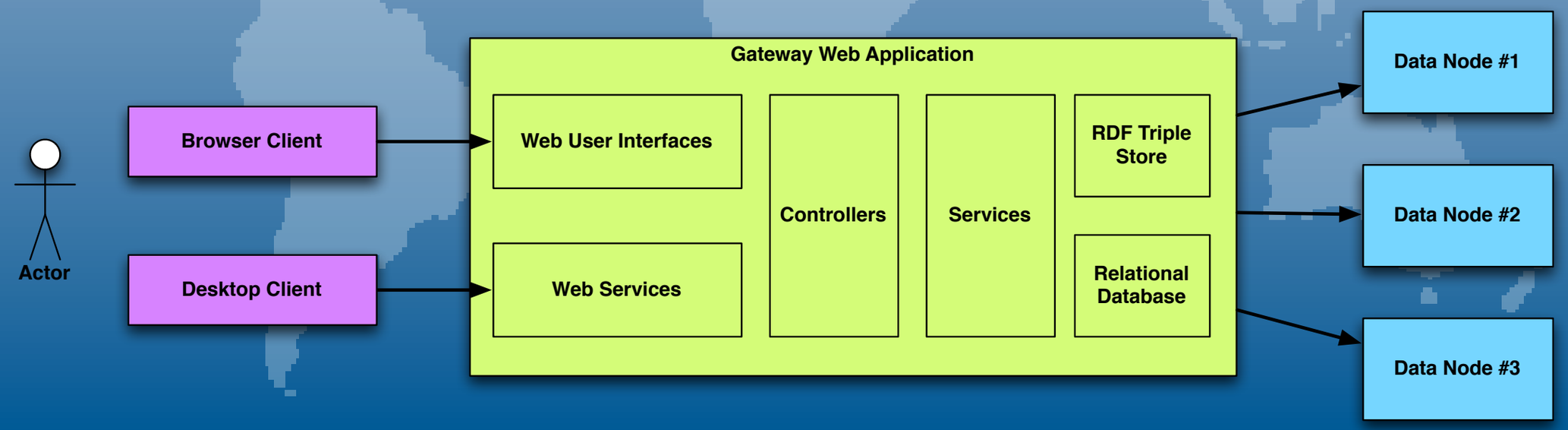


ESGF architecture: distributed, federated, scalable, integrated

ESGF Gateway

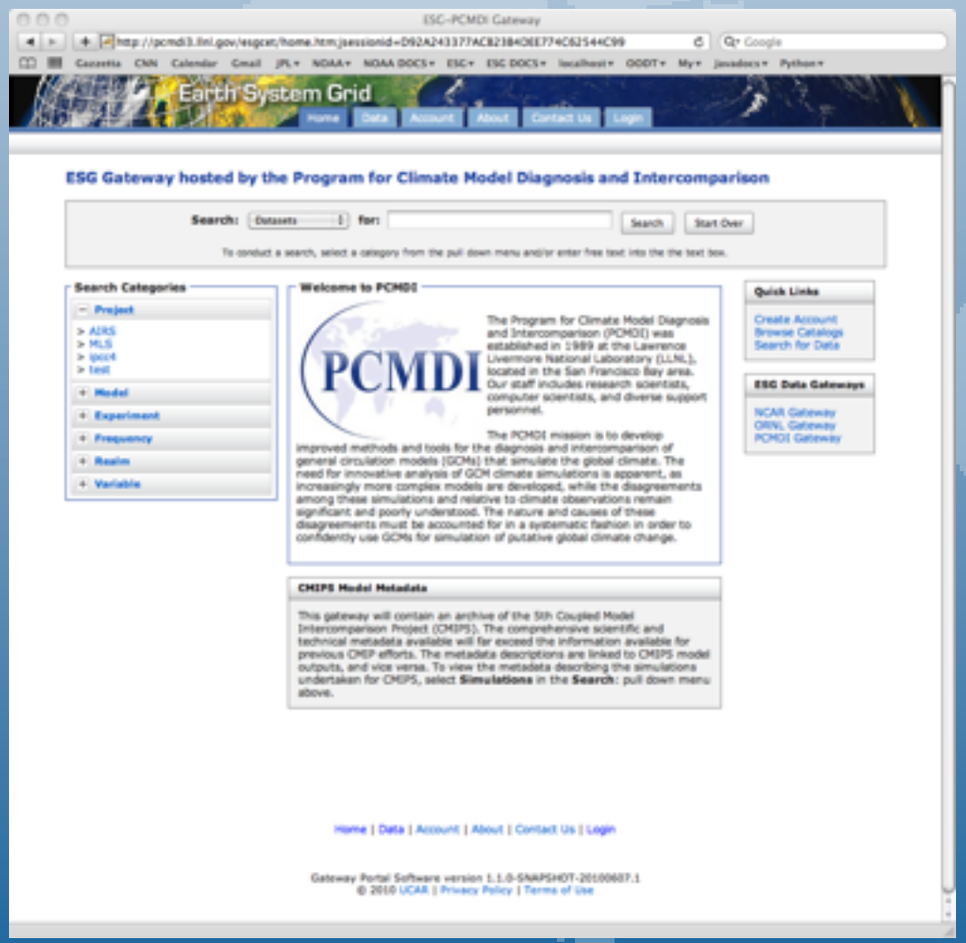
Web portal that enables and controls access to distributed data

- Entry point for user to discover, locate and request data products
- Operated autonomously by one institution to expose discipline-specific data holdings or data services
- Exposes interfaces for browser and rich-client access
- Index data from multiple data nodes
- Federates with other Gateways
- Functionality: user registration and management, authentication, authorization, metadata search, metadata display, files download, derived data product request



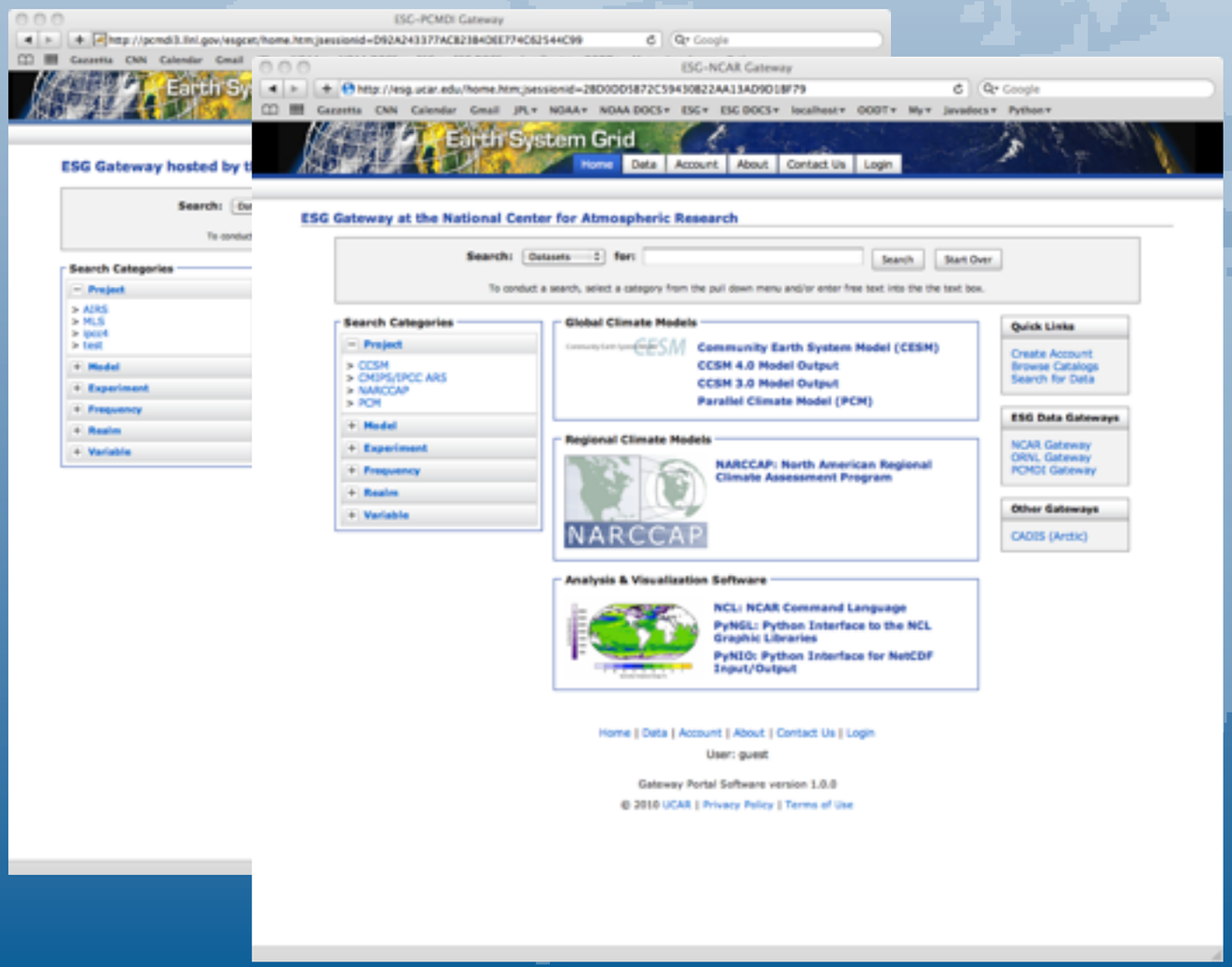
ESGF Gateways

- PCMDI Gateway: CMIP3 & CMIP5
- NCAR Gateway: CCSM and NARCCAP model output
- ORNL Gateway: observations (CDIAC, ARM, C-LAMP, …)
- NASA-JPL Gateway: satellite observations (AIRS, MLS, TES, …)
- Coming soon: BADC, DKRZ, Jamstec, ANU, NOAA NCMP & EPC



ESGF Gateways

- PCMDI Gateway: CMIP3 & CMIP5
- NCAR Gateway: CCSM and NARCCAP model output
- ORNL Gateway: observations (CDIAC, ARM, C-LAMP, …)
- NASA-JPL Gateway: satellite observations (AIRS, MLS, TES, …)
- Coming soon: BADC, DKRZ, Jamstec, ANU, NOAA NCMP & EPC

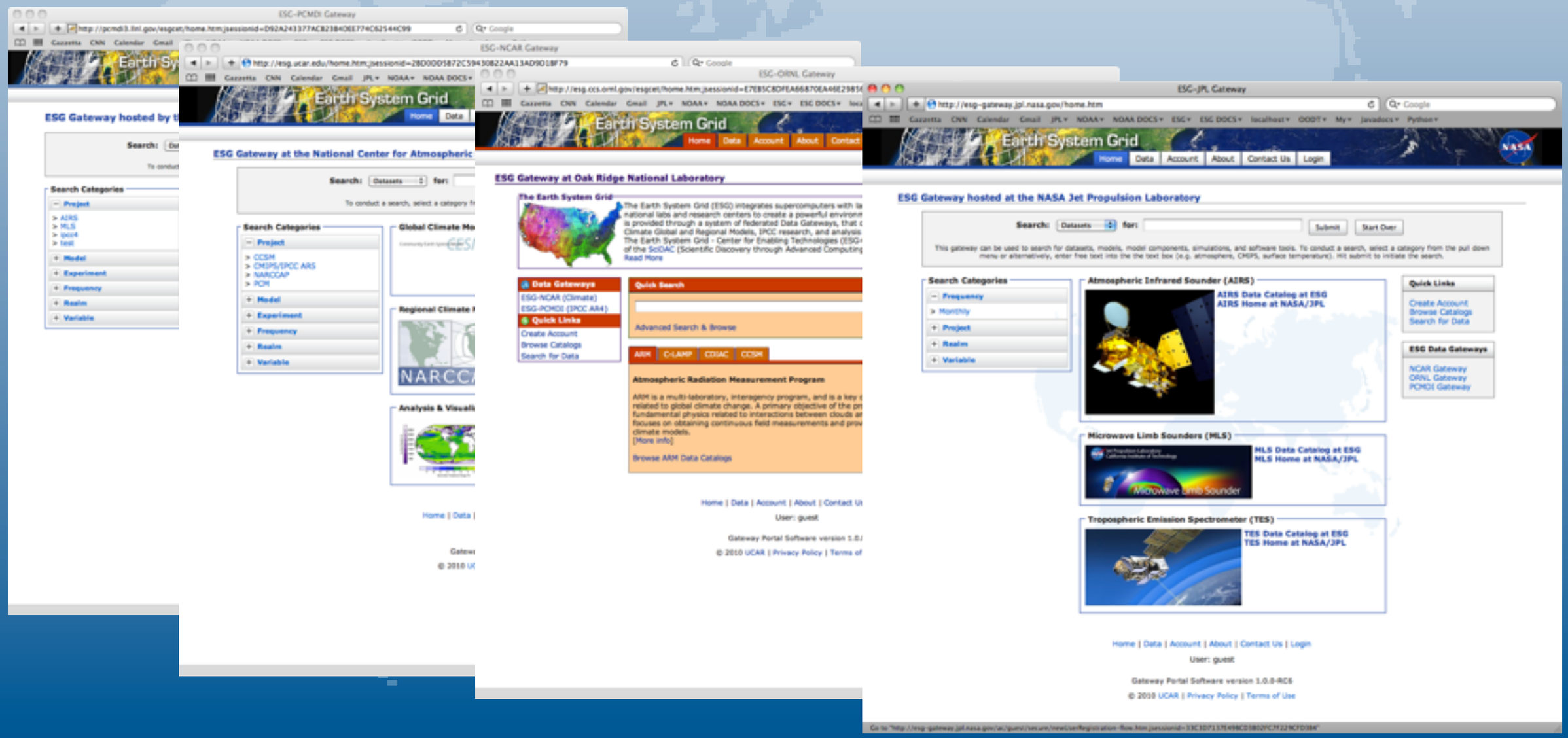


ESGF Gateways

- PCMDI Gateway: CMIP3 & CMIP5
- NCAR Gateway: CCSM and NARCCAP model output
- ORNL Gateway: observations (CDIAC, ARM, C-LAMP, …)
- NASA-JPL Gateway: satellite observations (AIRS, MLS, TES, …)
- Coming soon: BADC, DKRZ, Jamstec, ANU, NOAA NCMP & EPC

ESGF Gateways

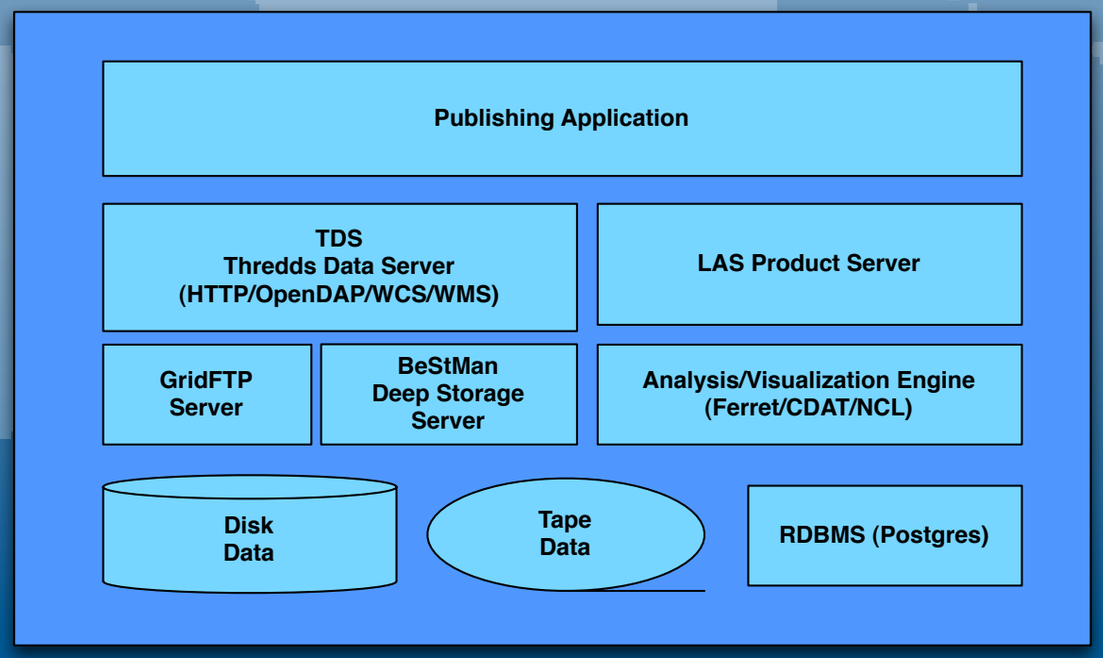
- PCMDI Gateway: CMIP3 & CMIP5
- NCAR Gateway: CCSM and NARCCAP model output
- ORNL Gateway: observations (CDIAC, ARM, C-LAMP, …)
- NASA-JPL Gateway: satellite observations (AIRS, MLS, TES, …)
- Coming soon: BADC, DKRZ, Jamstec, ANU, NOAA NCMP & EPC



Data Node

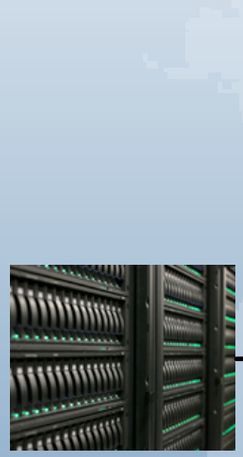
System (hardware+servers) where data is stored, published and accessed

- Data storage (online or deep archive)
- Data product services (files download, sub-setting, visualization ...)
  - Integrates common applications such as TDS, LAS, CDAT, ...
  - Flexible, custom configuration of services
- Data storage and processing are distributed and scalable
- Relatively easy entry point for data providers
- Expected approx. 20 sites for CMIP5 + others



Earth System Grid

Data Flow Topology



DATA NODE (Model Output)

Tape Data

Metadata Catalogs

Publishing Application

Data Servers

Data Provider

DATA NODE (Observational Data)

Metadata Catalogs

Disk Data

Publishing Application

Data Servers

X509

GATEWAY

MyProxy Server

Full Metadata RDBMS (Postgres)

Search Metadata RDF Triple Store (Sesame)

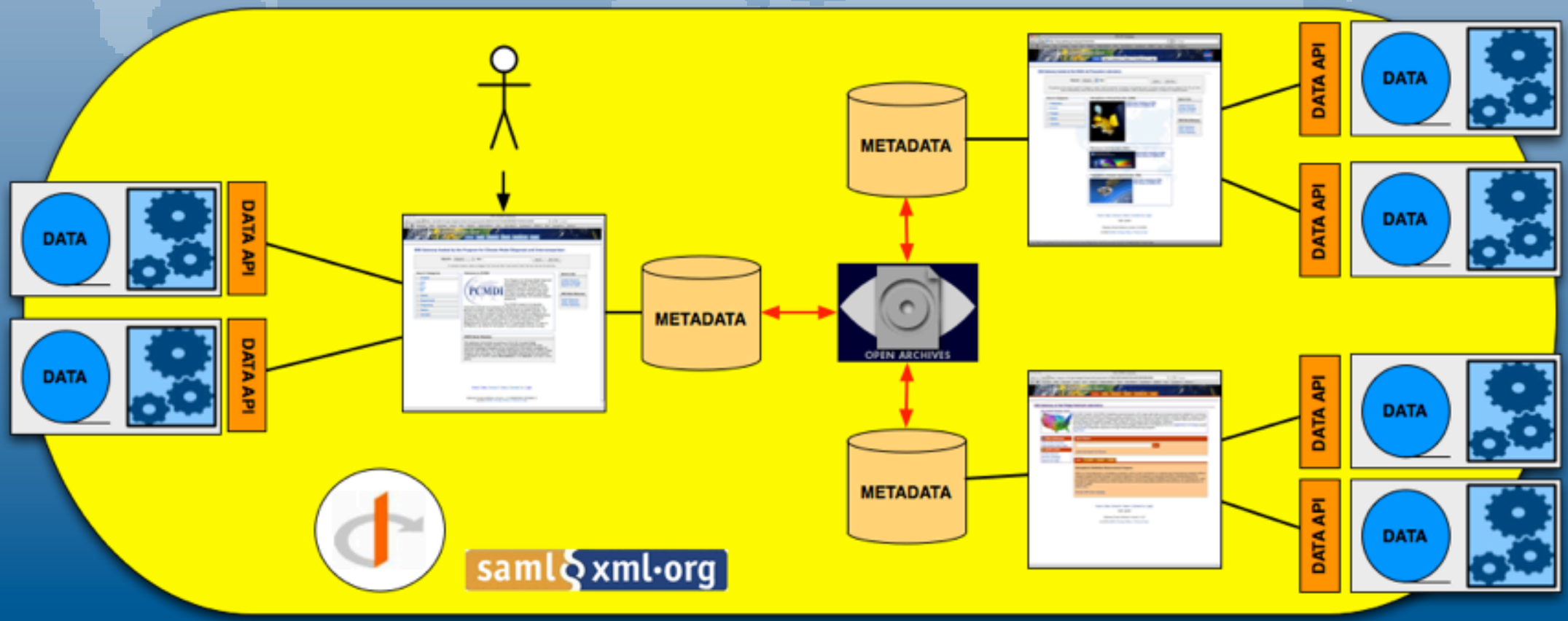
GATEWAY

User

User

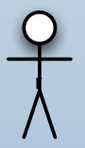
ESG Federated Services

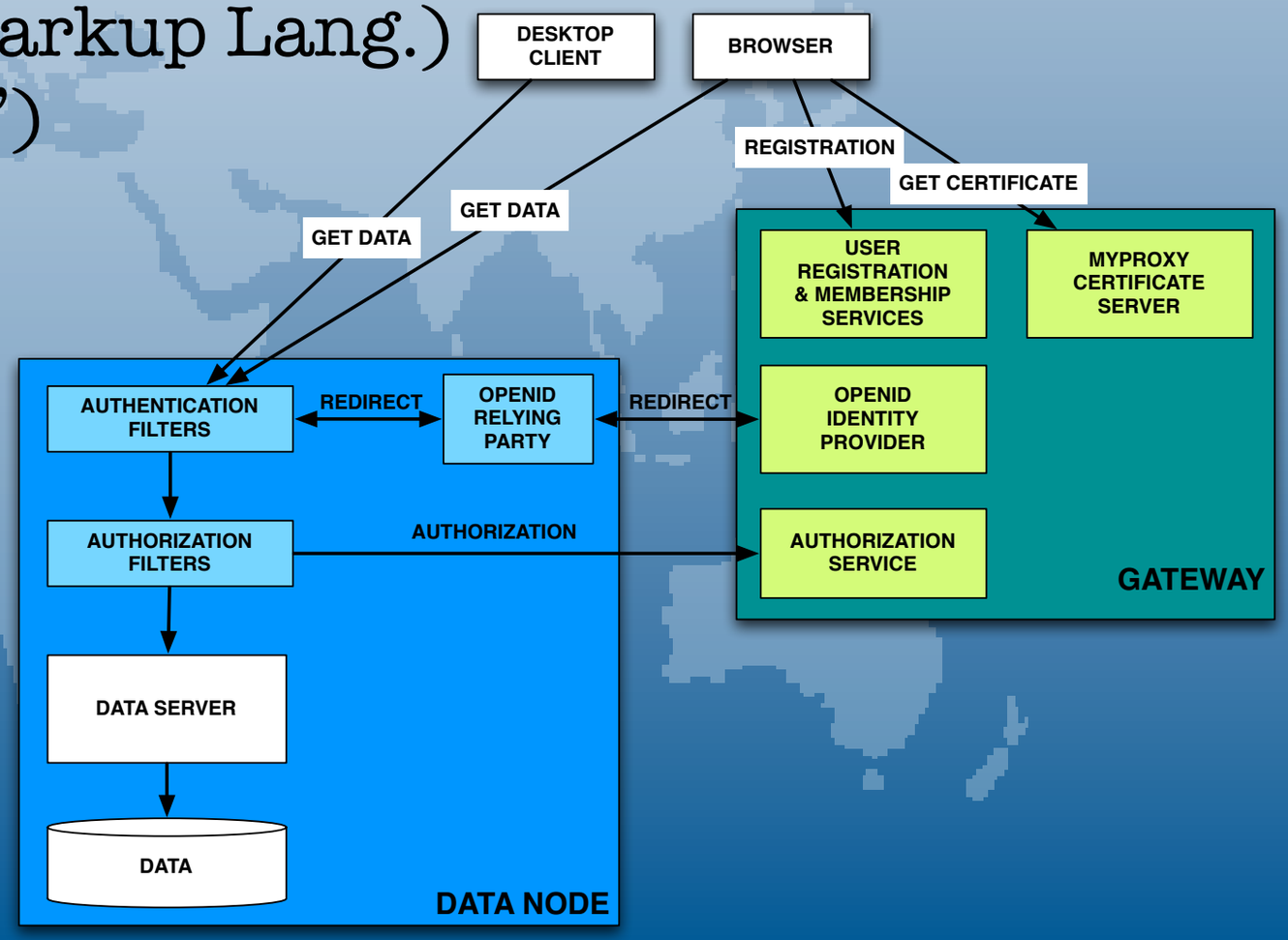
- Federation is a virtual trust relationship among independent management domains that have their own set of services
  - User authenticates once to gain access to data across multiple systems and organizations
- Federation is enabled by 3 key components:
  - Common security protocols (and trust relationships)
  - API and services for metadata exchange
  - Service-oriented data access model



ESG Security Infrastructure

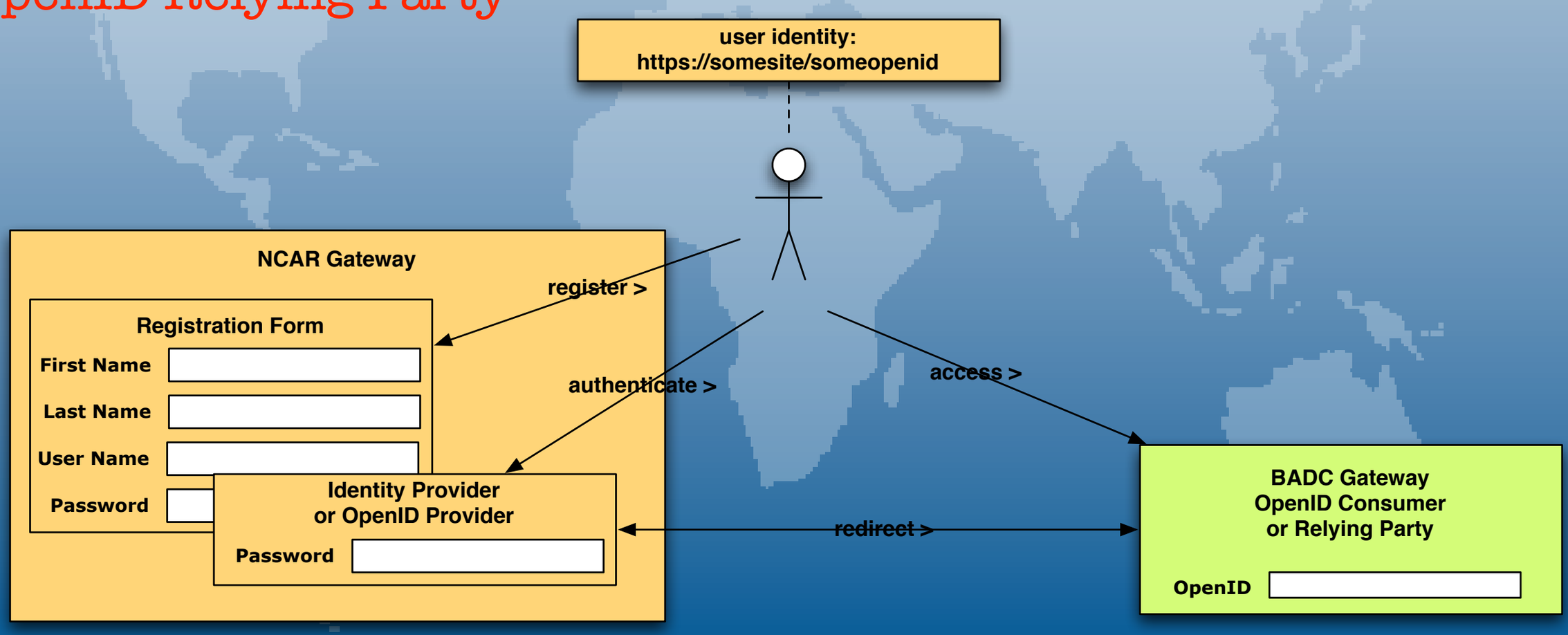
- Security infrastructure is used to restrict access to certain class of resources to authorized users only and to report usage metrics
- Must work for browser and rich client-based access
- Technologies:
  - OpenID
  - SAML (Security Assertion Markup Lang.)
  - X509 certificate (“enhanced”)
- Gateway components:
  - Registration services
  - OpenID IdP
  - Authorization services
  - MyProxy
- Data Node components:
  - Access control filters
  - OpenID Relying Party





OpenID

- Increasingly popular standard for web Single Sign On
  - Yahoo, AOL, ...
- User needs register at only one Gateway, and authenticate once during a working session, to be able to carry his authenticated identity throughout the federation
- ESG Gateway operates as both OpenID Identity Provider and OpenID Relying Party

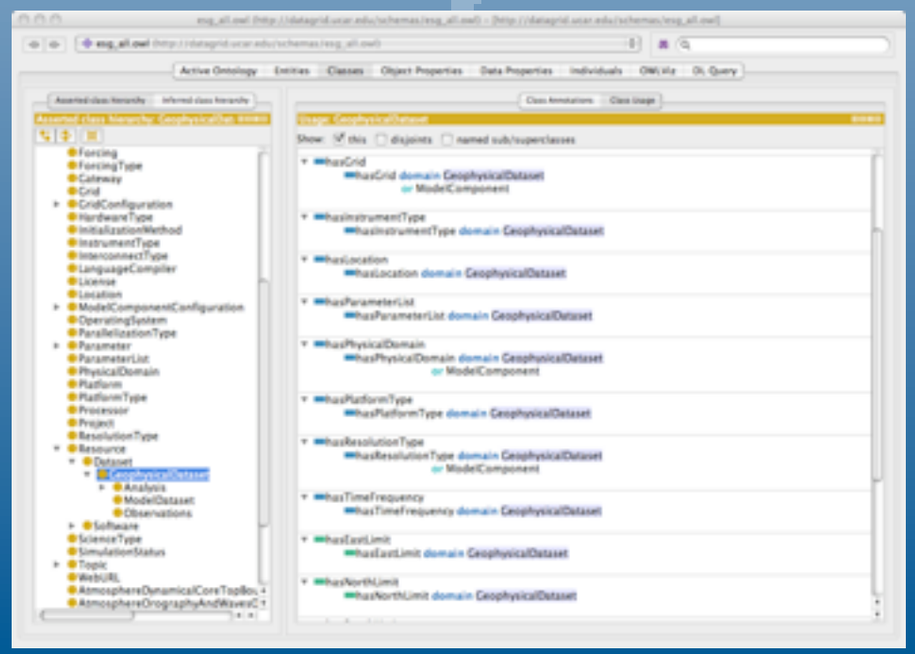


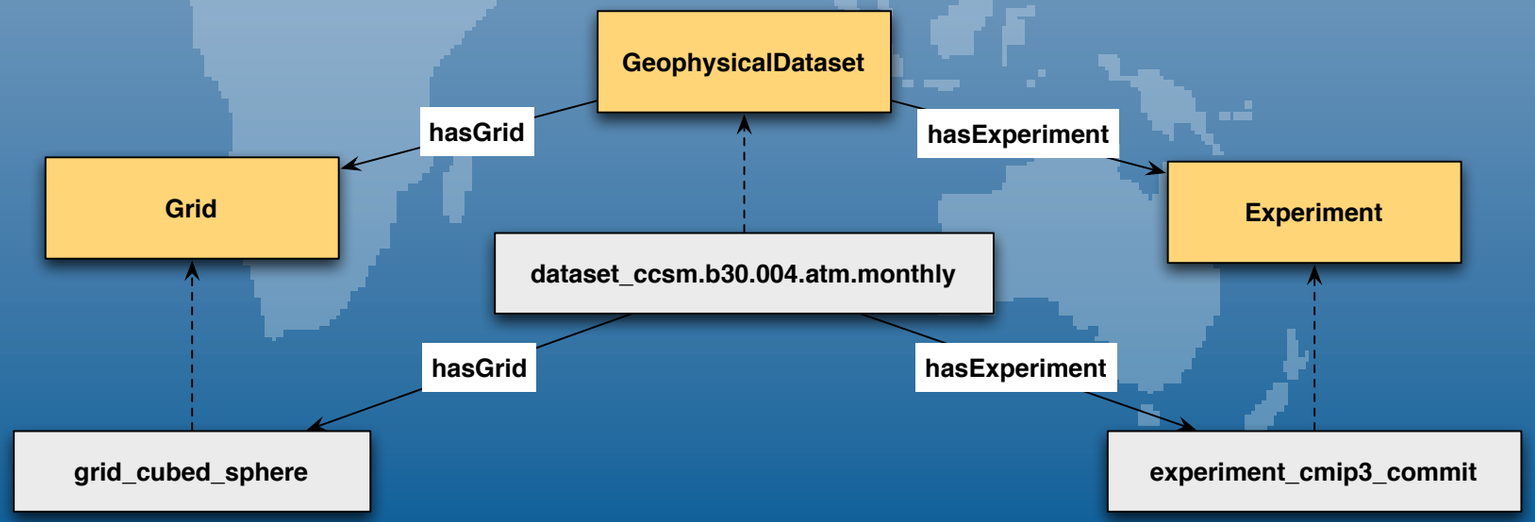
ESGF Cross-Federation Search Services

- Searching is primary tool for users to locate data of interest
- Gateway UI features combined “facets”+text search
- Search facets are dynamic and customizable
  - CMIP5 facets: model, experiment, frequency, variable, realm
- Returns results across ESG Federation
- Currently backed up by metadata stored in a semantic repository

ESGF Semantic Infrastructure

- Search services implementation based on semantic technologies such as RDF, OWL and Sesame Triple Store
- ESGF OWL ontology contains classes, properties and individuals
  - Classes: Dataset, Instrument, Grid
  - Properties: hasGrid, hasGeoLocation…
  - Individuals: ccsm_dataset_1, tripolar_grid
- Facet constraints follow semantic relations between objects
- All metadata harvested from Data Nodes is converted to RDF triples (conforming to the OWL ESGF ontology) and stored in a Sesame triple store





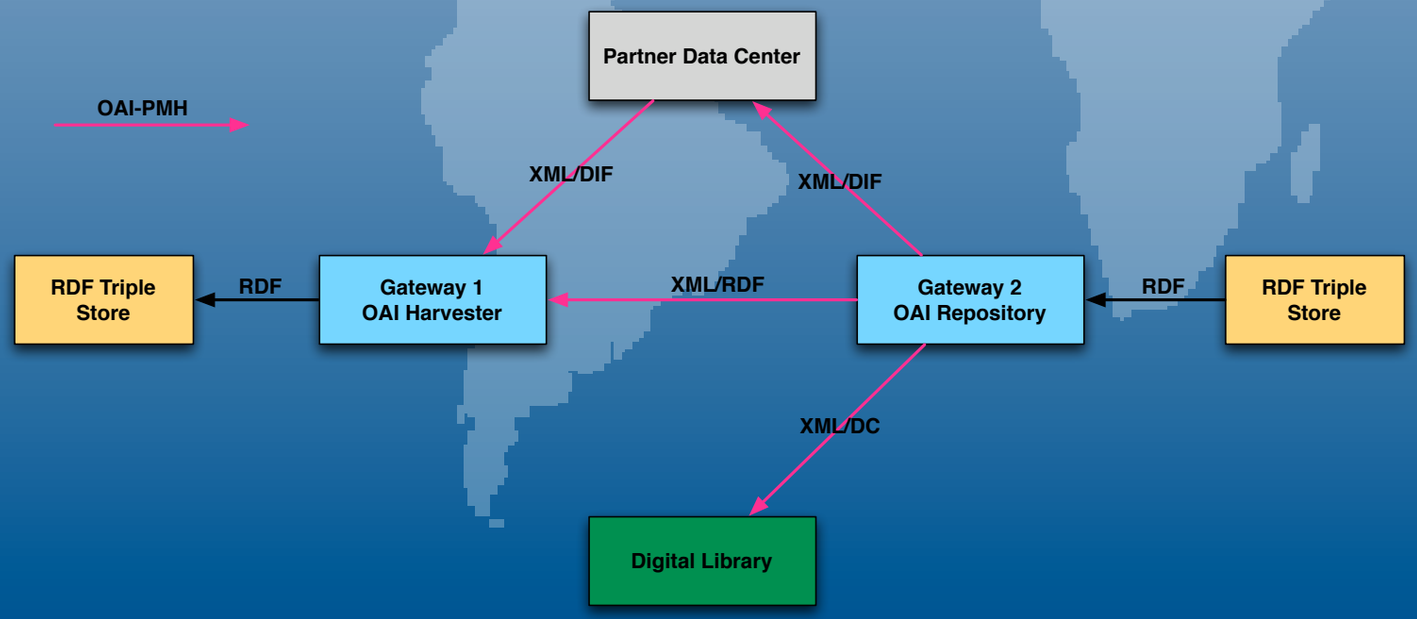
Detailed Semantic Model Metadata

- Modeling centers to enter detailed model metadata into Metafor Questionnaire web application
- ESG Gateway @PCMDI to capture metadata via Atom feed
- Metadata converted to RDF/OWL and hyperlinked to model search results for unprecedented in-depth access to model/simulation configuration and comparison

OAI-PMH Metadata Exchange

Open Archive Initiative–Protocol for Metadata Harvesting

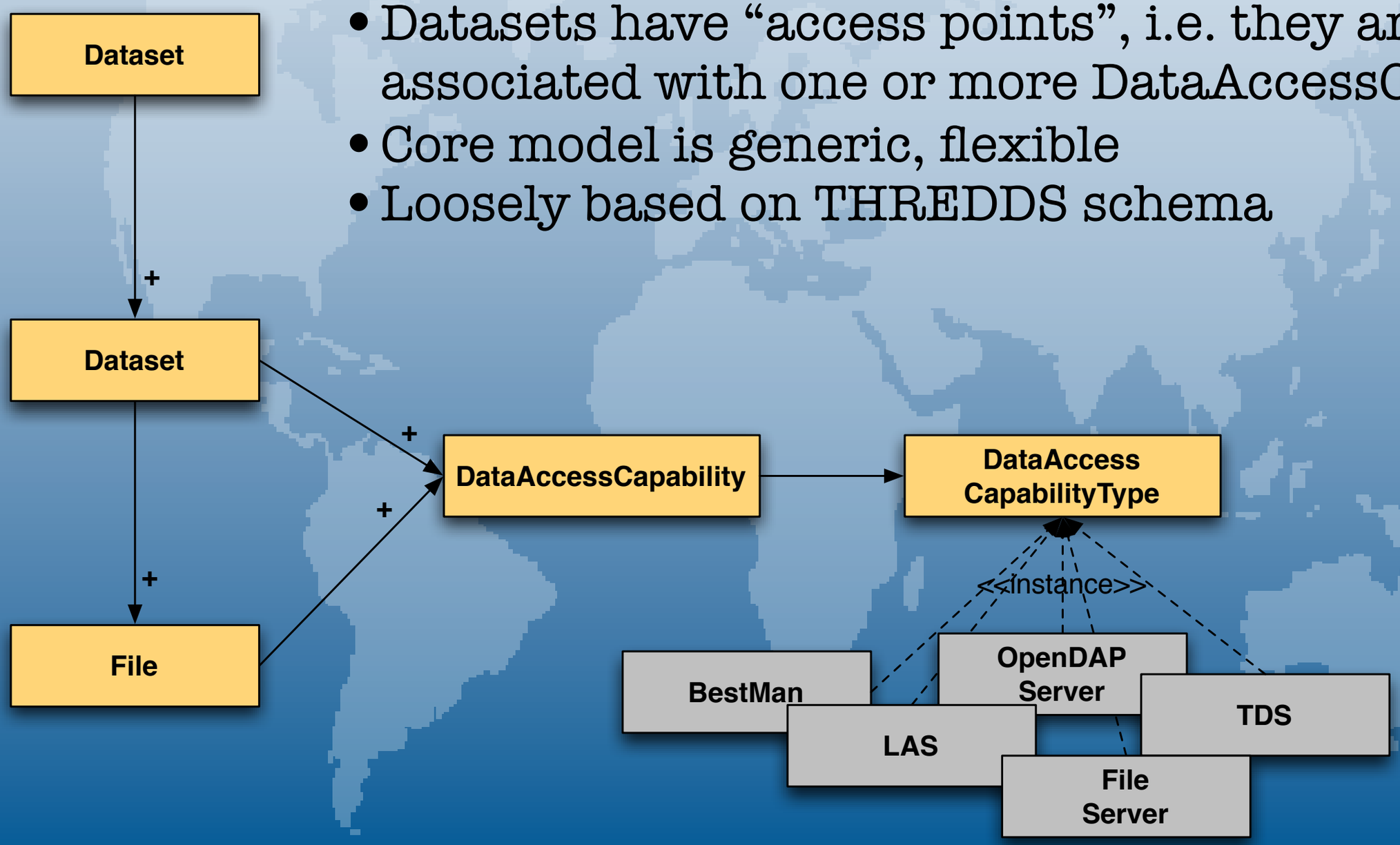
- Client (OAI Harvester) – Server (OAI Repository) protocol
- Metadata records are encoded in XML format (any schema)
- 6 "verbs": Identify, ListMetadataFormats, ListSets, ListIdentifiers, ListRecords, GetRecord
- Widespread use among digital libraries, extending to geo-scientific domain (IPY, WMO) and planetary science
- Each ESGF Gateway can act as an OAI Repository and Harvester
  - Import/export records into/from RDF
  - Supported metadata formats: RDF, DC, DIF, etc.



ESGF: a global network for the exchange of semantic information

Service-Oriented Data Access Model

- Dataset as data “container”
- Datasets are organized hierarchically
- Datasets have “access points”, i.e. they are associated with one or more DataAccessCapability
- Core model is generic, flexible
- Loosely based on THREDDS schema

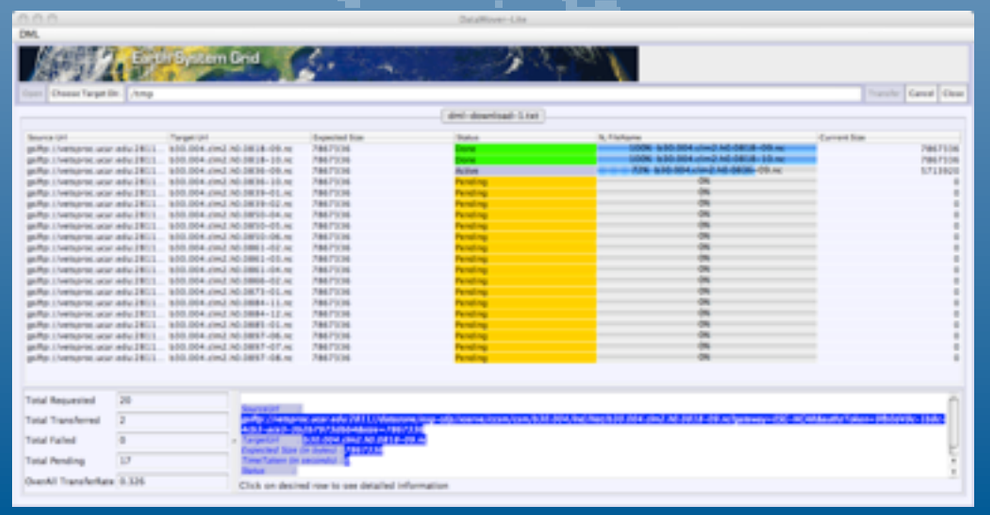


Example: Distributed Data Download

User can download files from distributed sources in 3 ways:

1. Direct HTTP hyperlinks
   - Browse datasets on multiple Gateways
   - Click on individual hyperlinks
   - Single Sign On through OpenID

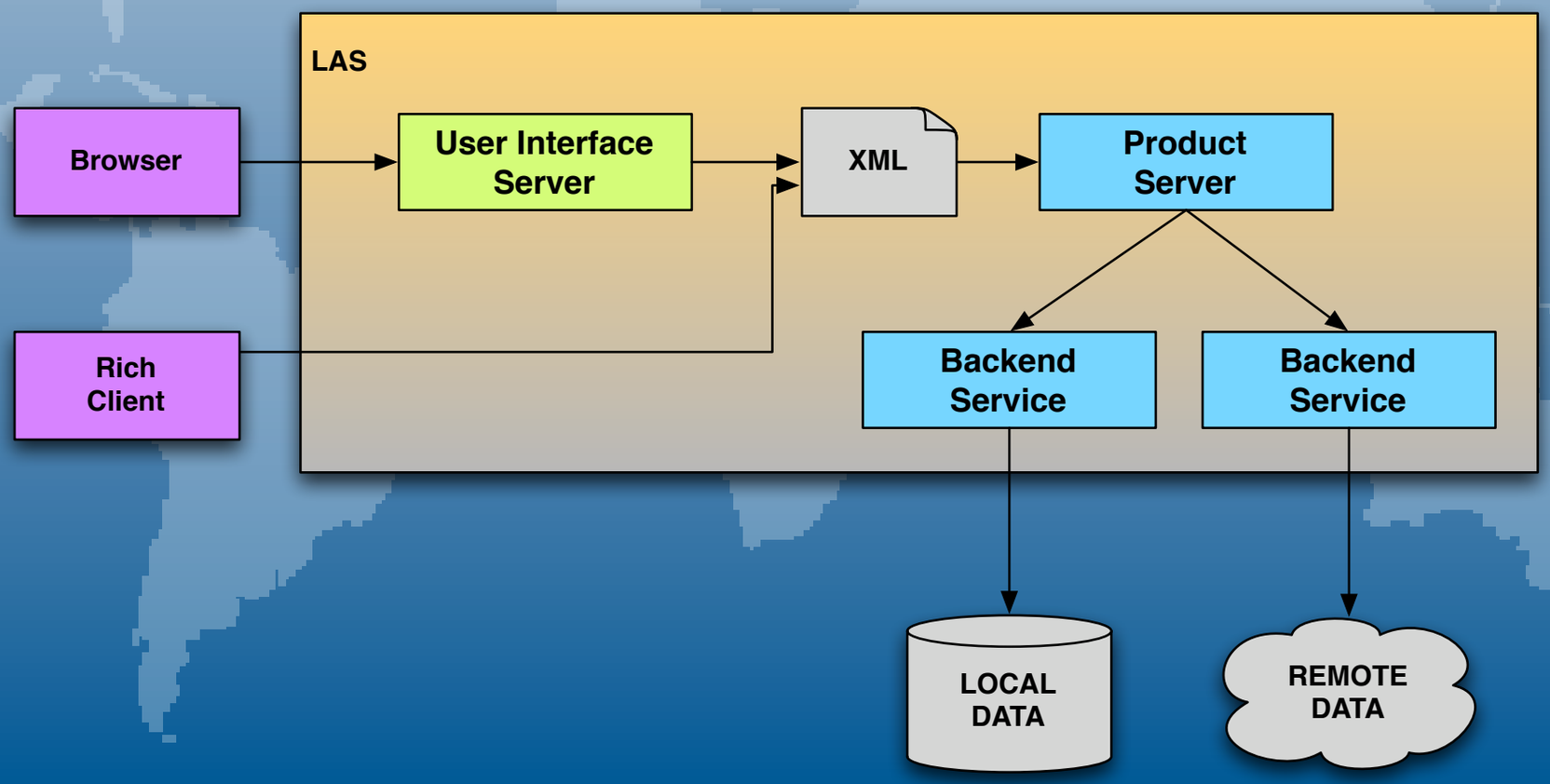
2. wget/curl script
   - HTTP bulk download of files
   - Authentication via X509 certificate



3. Data Mover Light (DML)
   - Rich desktop client with GUI
   - Downloaded via web-start
   - Authentication via X509 certificate
   - Data transfer via GridFTP

Distributed Data Analysis & Visualization via LAS

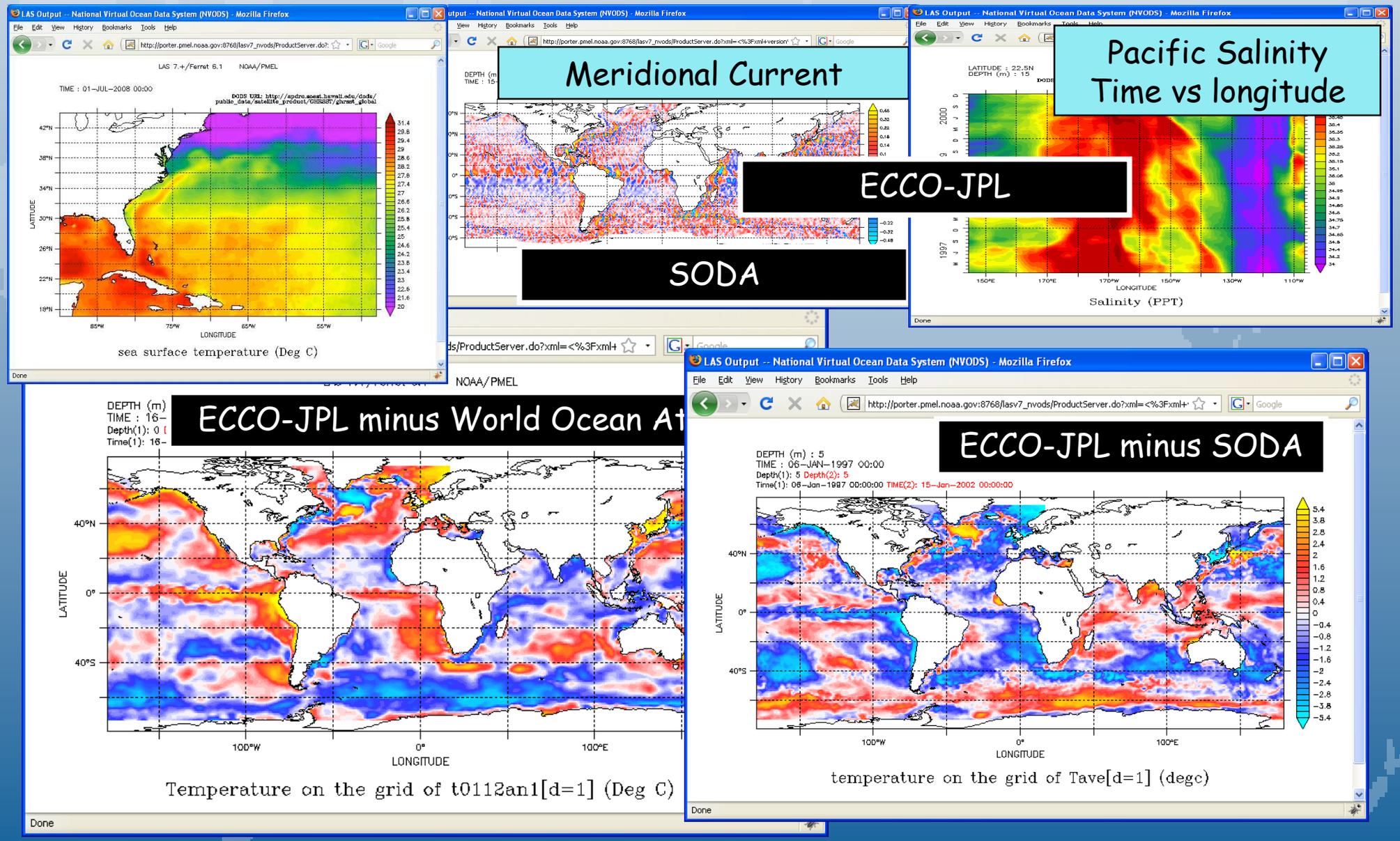
- Live Access Server (LAS): Analysis & Visualization engine developed by NOAA/PMEL
- Access and analysis of data from distributed sources via OpenDAP
- Integrated as default data product server within ESG architecture
  - May be exposed as a DataAccessCapability on a Data Node
  - Working on integration with ESG security (auth/authz)



Earth System Grid

Live Access Server (LAS)

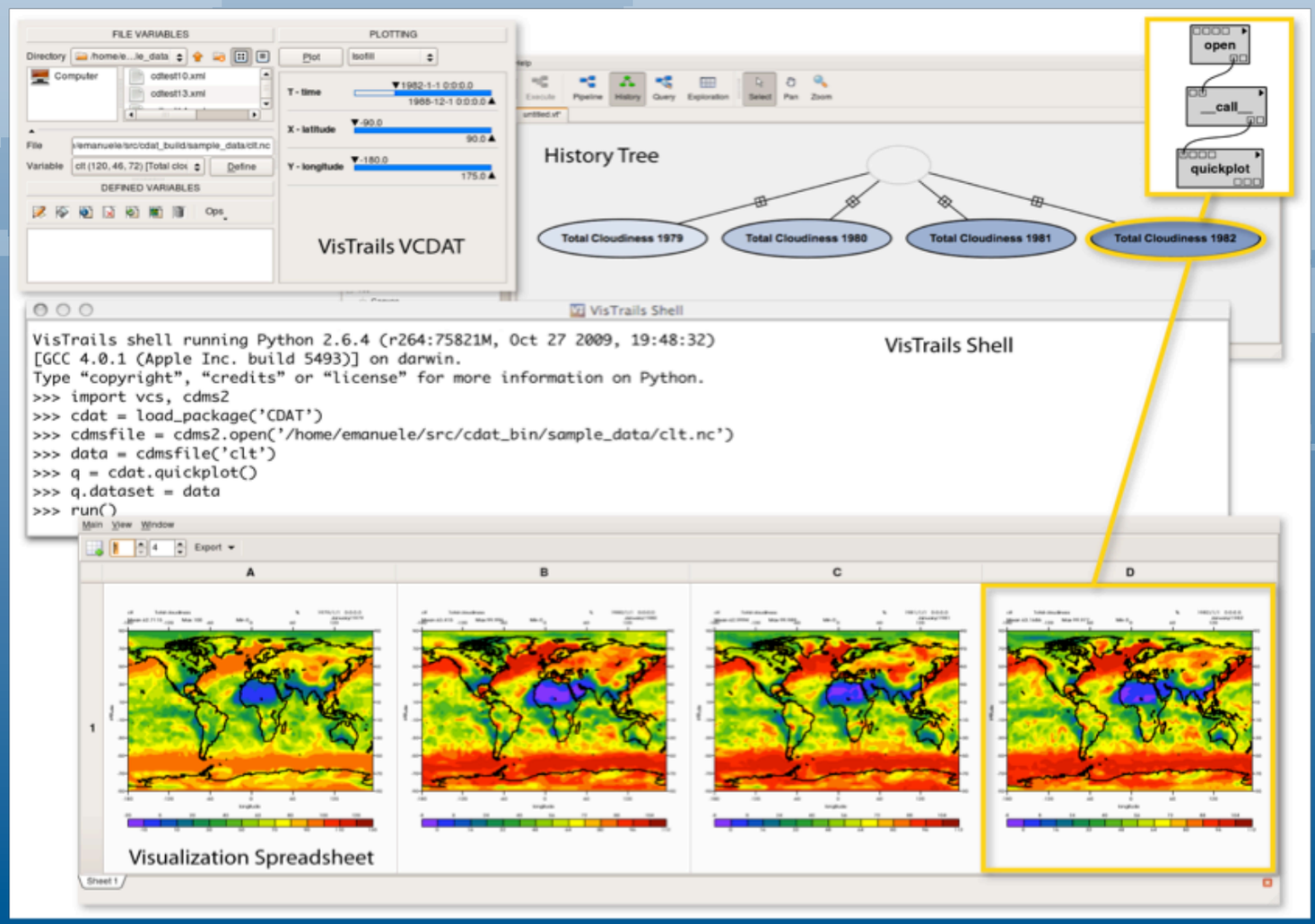
Example: multiple distributed datasets (model outputs, satellite feeds and climatologies) are retrieved, re-gridded and compared.



Courtesy of NVODS (National Virtual Ocean Data System)

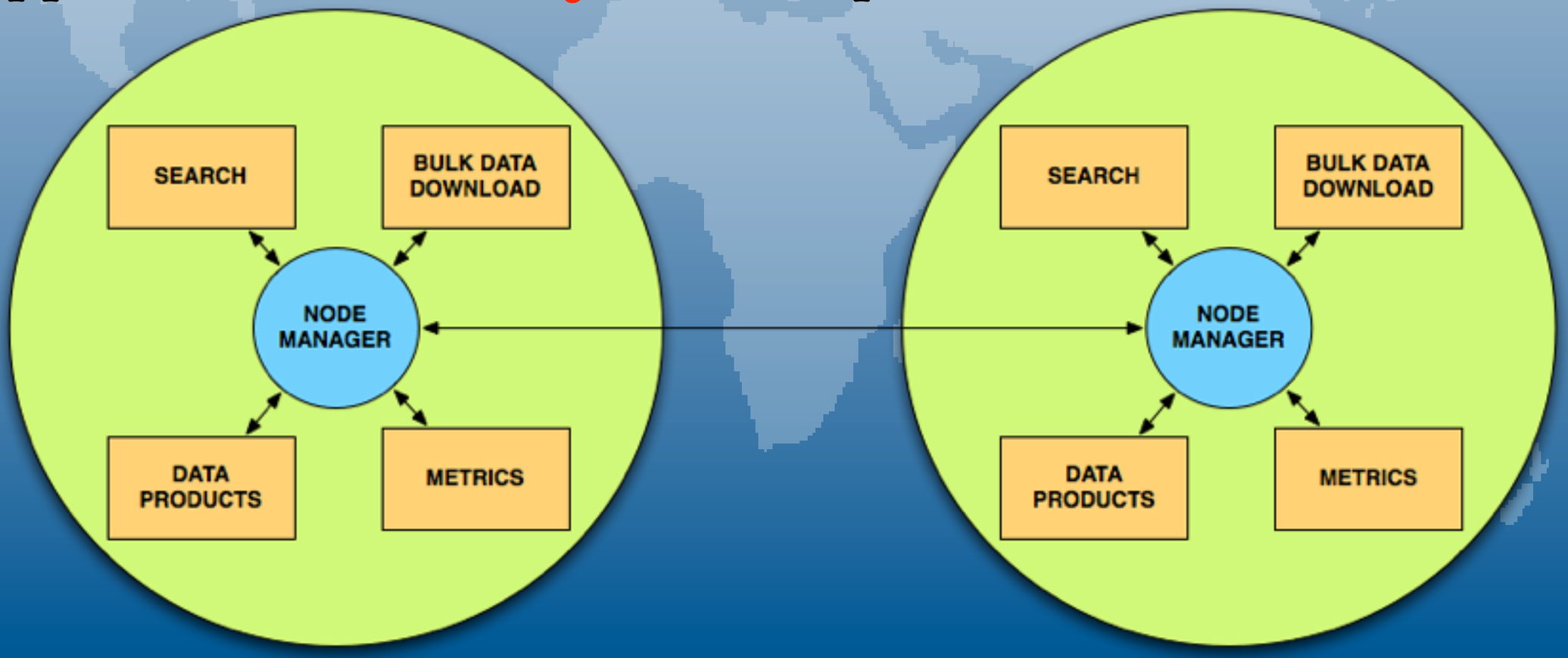
Distributed Data Access via Rich Client

- UV-CDAT: Ultra-scale Visualization Climate Data Analysis Tools
- Open source BSD analysis software consortium: PCMDI, ORNL, LBNL, LANL, University of Utah/SCI, Kitware, Tech-X, BADC, ANU
- Integrates of several software packages: CDAT, VisTrails, etc.
- Includes workflows and provenance
- Picture: data access from 4 Data Nodes



Next Generation ESGF Architecture

- Evolving towards a peer-to-peer architecture
  - Generic "Node" that replaces a Gateway Node or Data Node with totally customizable functionality
  - Loosely coupled components interacting via message events
    - Subscription-Notification mechanism
  - Allow for multiple swappable implementations of same component functionality
- Support for Java and Python components



New Development Paradigm

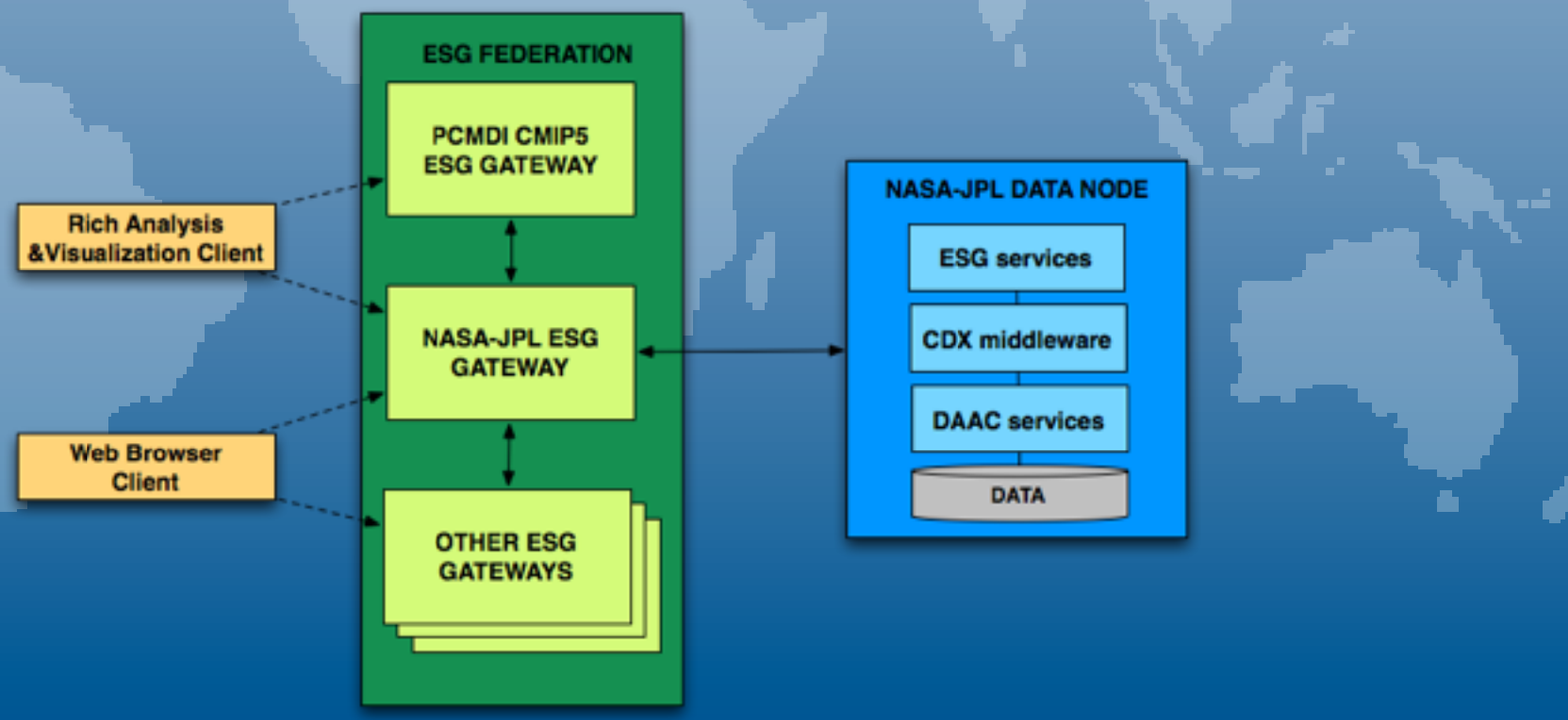
- All software to have Open Source license (BSD)
  - can be downloaded, modified and redistributed
- Community Driven development
  - Accept contribution from many institutions
  - Published APIs that can be used to develop custom components or implementations
- Collaborative development tools:
  - Common distributed software repositories
  - Dependency management
  - Periodic integration builds
  - Shared centralized documentation
- Governance
  - Governance and priorities provided by GO-ESSP (PIs)
  - Technical Working Group provides architectural guidance and technical coordination
  - Light cross-federation project management to keep track of milestones and deadlines

Inclusion of Observations

- Goal: include observational datasets into CMIP5 archive to:
  - Score models when evaluating ensemble predictions
  - Improve understanding of physical processes and model prediction capabilities
- Challenges:
  - Heterogenous data sources (satellite, ground stations, field missions)
  - Heterogenous data formats (HDF, HDF-EOS, NetCDF, ASCII,...)
  - Incomplete and non-homogenous metadata
  - ESG metadata model designed for model data (no concept of “instrument”, “platform”, etc.)
  - Data already stored on legacy systems
- Strategy: structure the observations as close as possible to model data to facilitate use and comparison
  - Same data format (NetCDF)
  - Same metadata conventions (CF)
  - Same hierarchical organization (DRS)
  - Include “caveat” documents containing usage instructions, description of uncertainty etc.
  - Make observations and model output discoverable via same faceted search
- Interested parties: NASA, NOAA, ORNL, ...

IPCC Strategy at NASA-JPL for next 18 months

- Already deployed and operating a prototype ESG-NASA-JPL sub-system composed of 1 Gateway + 1 Data Node
  - Next: more datasets, services, documentation, & CMIP5 integration
- Proposed NASA "science" workshop to select mission data most critical to IPCC, adopt same standards for data/metadata formats and for documentation
- Proposed NASA "technical" workshop to coordinate software infrastructure necessary to support the IPCC/AR5 science objectives
- Integration with PO.DAAC based on combination of existing DAAC services, ESG infrastructure and CDX middleware



NOAA-ESRL IPCC Involvement for next 18 months

- Environmental Projection Center (EPC)
  - Provide tools and services to enable use of climate downscaled data for regional decision making
  - Scope: regional (100km) to local (1km) space scale, annual to decadal time scale
  - Prototype under design at ESRL, starting with use case definition and requirement gathering
- National Climate Model Portal (NCMP)
  - Developed at NCDC, will start new collaboration with ESRL
  - Three audiences:
    - Education for K-12 schools
    - Regional impact community (agriculture and water)
    - Climate modelers
  - Integration of NOMADS and ESG technologies
  - Clearing house for re-analysis data
- GIP: Global Interoperability Program
  - Coordination and funding of interoperability projects in climate and weather research, operational forecasting, climate change assessment
    - Document software infrastructure for Earth Sciences on Wikipedia
    - Sponsor summer workshops and colloquiums

Summary

- Earth System Grid Federation (ESGF): open, spontaneous collaboration of international partners committed to the development of a global technology infrastructure for Climate Change research
- ESGF Architecture: distributed system of Gateways and Data Nodes bound together by Federation Services:
  - Security services based on standard technologies, and trust relationship among partners
  - Common protocols for metadata exchange
  - Flexible service-oriented data access model
- ESGF major current thrusts:
  - CMIP5/IPCC-AR5 support
  - Inclusion of observations from NASA, NOAA, ORNL, ...
  - Movement towards open source, community-driven software development and project governance